DOCUMENT RESUME

ED 296 737 IR 052 401

AUTHOR Kahn, Paul
TITLE Outline for Research in Large Data Base Resources.
INSTITUTION Brown Univ., Providence, RI. Inst. for Research in Information and Scholarship.
PUB DATE Apr 87
NOTE 15p.
PUB TYPE Reports - Descriptive (141)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Access to Information; Bibliographic Coupling; Citations (References); *Database Management Systems; *Databases; *Information Retrieval; Man Machine Systems; Online Searching; *Research Tools; *Search Strategies; Users (Information)

ABSTRACT

This paper uses a hypothetical application entitled "VAPORTRAILS" to examine how an integrated application can be used to solve the problems of search and retrieval from a range of qualitatively different databases, and the organization of the resulting information into a personal database resource. In addition, four general classes of databases are described which are important to the scholar's work environment. The use of "Intermedia," a hypertext system under development at the Brown University Institute for Research and Scholarship to address some of these integration problems, is then considered. An environment which includes the current Intermedia applications (InterText, InterDraw, InterVal, and InterSpect) and the overall Intermedia linking mechanism is described. Several extensions are added, including VAPORTRAILS and a new kind of link that will use search or look-up applications from a selected point in one document to the appropriate information in another data collection. Ways of searching in an Intermedia web are suggested, including the ability to access block information in the Intermedia documents in a manner analogous to the system used by VAPORTRAILS to access and manage citation information. One figure is provided. (Author/EW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Outline for Research in Large Data Base Resources

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paul Kahn
Institute for Research in Information and Scholarship (IRIS)
Brown University
April 1987

ED296737

Abstract

The first part of this paper gives an overview of the place large data bases now have in the scholar's computing environment. Some of the current technological developments which are affecting this part of computing are mentioned, and the current limitations are highlighted.

The next part of this paper is a description of a hypothetical application to which we have given the name VAPORTRAILS (to underscore its current non-existence). This description is an attempt to flesh out what we mean by an integrated application's solution to the problems of search and retrieval from a range of qualitatively different data bases, and the organization of the resulting information into a personal data base resource. VAPORTRAILS is intended to provide the tools of the information specialist directly to the end user. Important parts of this description are built on the Graphical Icon-driven Search Tool (GIST) prototype.

The next part of this paper describes the four general classes of data bases which are important to the scholar's work environment. We then go on to explore how Intermedia, the hypertext system under development at IRIS, could be extended to address some of these integration problems. An environment which includes the current Intermedia applications (InterText, InterDraw, InterVal, InterSpect) and the overall Intermedia linking mechanism is described, to which we add several extensions. One of these is the VAPORTRAILS application itself. In addition, a new kind of link which will use search or look-up applications for following from a selected point in one document to the appropriate information in another data collection is described. Ways of searching in an Intermedia web are suggested. We add to this the ability to access Block information in Intermedia documents in a manner analogous to how VAPORTRAILS accesses and manages citation information.

1. Introduction

At the highest level of abstraction the scholar is an information worker. The scholar's information base, the stuff or material with which he or she must interact, can be viewed as organized into various collections. The books in a university library, the specimens in a museum collection, the field notes of a scientific expedition, and the statistical reports of government bureaus are all common examples of non-electronic data collections. These data collections can be described in terms of their general characteristics. Many kinds of data collections exist in a scholar's working environment. A key practical question is: of these collections, which are appropriate for delivery in electronic form? An even more important question is: given that some can or will exist in electronic form, what kinds of access tools are required in order to make such electronic data collection as useful as possible? These are the overriding questions which we are addressing in the discussions which follow.

1.1. Traditional Large Data Bases

A number of forces come into play in the creation of and access to large data bases. Traditionally this is the field of information science. On the technical side, the driving forces are in data collection and manipulation: data processing. In the same way that our models of physical libraries are driven by the physical need to store, and later access, books on shelves, our ways of collecting data have been driven by our models of how to store information on machines and how those machines manipulate the information bits. If the amount of information to be stored is large, the machine required to store it must have large storage space and must have powerful data base management software. Since the inception of using computing machinary to manage collections of information *large* has been equal to *expensive*. Dedicated staff are required to maintain such an expensive machine, and since their time is dedicated to this machine no

IR053401

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Nicole Yankelovich

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

software interface is too complex for their attention. Each data collection has been built upon its own machine/software base. As in the example of the data processing department of a large corporation, if someone wants information from the data base, they ask the keepers of the machine for a report. The keepers of the machine then generated a report from the data base and delivered this to the end user.

As remote access to either batch or interactive computer systems became possible a new class of intermediary came into play: the *information specialist*. It is here that library science entered the field. This person could now have access to a number of different data collections, each of which was qualitatively different in some manner. This person's job was to understand how to interact with the variety of interfaces presented by different data base management systems.The end user would come to the information specialist seeking the answer to a question. The information specialist would act as the reference librarian for various electronic data collections, but with one important difference. Where the reference librarian could tell the end user what printed sources to look at and leave the searching to the user, in the case of electronic data collections the information specialist was the only person with the training and equipment to do the search itself. This specialist would access one or more remote computer systems, do a search, print the results, and deliver them to the end user.

To facilitate greater access to data collections over the growing telecommunications networks collections of data collections have been created such as Dialog, BRS, VU/TEXT, and ORBIT. The *on-line industry* came into being. In these cases the fundamental data base management system has been unified, though the implementation in individual data bases still varies in important ways. There might now be one syntax used to search for news stories in thirty-five different newspaper data bases on VU/TEXT, but the way in which stories about *U.S involvement in Nicuragua* or *French atomic testing in the South Pacific* are indexed will still vary from newspaper to newspaper.

1.2. The Advent of Personal Access

The addition of personal computers in the workspace of both the information specialist and the end user is now having an impact on the way large data bases are stored and accessed. Just as the addition of a remote terminal with telecommunications gave rise to the information specialist and the on-line industry, the personal computer has given rise to a new class of applications software designed to bring the information closer to the information specialist or to the end user's own machine, the personal computer. A number of communication software packages have been developed to make the mechanics of "calling up" a remote data collection as simple as possible. On the personal computer itself applications have been developed to provide a general interface for accessing qualitatively different data bases. The first of these, In-Search (Menlo Corporation), was designed specifically for the peculiarities of DIALOG, but later additions to the market, such as Pro-Search (Menlo Corporation/Personal Bibliographic Software), and Sci-Mate Searcher (Institute for Scientific Information) are more generalized. On the other end of the telephone, data providers or network hubs such as BRS or EZNET have created menu front-end programs to help the user move through the *stacks* of data collections available on-line.

At the same time data base applications specifically designed for personal computers have become an industry in and of themselves. Most of these applications mimic the data processing applications which were designed for the large, complex machine model. In many cases they are simply an implementation of mainframe software, such as BRS/Search, MicroBASIS, and Oracle, recompiled on the Itel 80xx or 80386 microprocessor. But there has also grown a new form of the data base management system, the *personal* data base designed to manage bibliographic information. One such example, Professional Bibliographic System for the Macintosh and Pro-Cite for the IBM PC (Personal Bibliographic Software), offers additional software for making use of information gathered from large data base collection such as DIALOG, BRS, and NLM.

In the current personal computer software market these various applications functions (communications front-end, general interface to different data bases, personal data base management systems) are still viewed as discrete functions. No one has yet attempted to provide an integrated applications solution for dealing with a full range of data base resources.

1.3. Limitations of Current Facilities

With the exception of the personal data base applications, which are discontinuous with large on-line data collections, all of the current advances are basically limited by the same machine model. The lowest common denominator for a remote connection to a data base management system is the command line of a teletype computer terminal. Menuing systems may offer some convenience over the basic command-line interface from which all these systems start, but they do not solve important problems. No menuing system can provide enough information on a single 24 column by 80 row screen to adequetly guide the user through a complex query. Almost all data collections, such as Medline, Compendex, MLA Index, rely on large amounts of printed documentation and special training to provide the user with the kind of information they need to do a complex search. No one has found effective ways to make such information available to the user on-line.

The advent of optical storage media in the personal computer marketplace has sent cracks through the current boundaries between large remote and local personal data collections. With CD ROM providing a means to deliver a local copy of a data collection the size of ERIC to a personal computer user, the question now arises as to whether a microcomputer copy of mainframe software is desirable. Many of the products coming onto the CD ROM market in 1986/1987 struggle with this question. The notion of integration of data with other applications is well illustrated in the two major CD ROM information products available in the business market from Datext and Lotus Development Corporation. In both cases collections of textual and numerical data about publically traded companies are stored on CD ROM, and can easily be copied from the data base access application into word processing and spreadsheet applications.

Engineering workstation and personal computer software have now developed *windowing* interfaces which can use graphic icons to represent programs and documents. Such systems can display multiple kinds of information to the user on one screen, and where these run under multi-tasking operating systems each window can contain a discrete active application program. Such capabilities, either single-tasking or multi-tasking, are not yet to be found in the large data base world. Yet the CD ROM now provides one means through which large data bases can be *delivered* to these systems in an inexpensive manner.

In the sections which follow we offer an outline of how we would like to move toward solutions to these problems in the field of information for scholars.

2. A Description Of VAPORTRAILS

2.1. Definitions

I am interested in pursuing an **Entity** about which there exists some information. This Entity could be described variously as an idea, topic, concept, set of data, or subject of study.

I expect that this information is to be found somewhere in the collection of qualitatively different data bases available on the **System**. The System consists of any and all information collections which can be accessed directly or indirectly from the computer on which I work.

Such collections would include any information 1) on my local computer, 2) on the network of computers in my department, 3) over a campus network service such as BRUNET which provides access to resources in other departments and campus information servers, or 4) from a national network service such as Telenet which provides access to collections of qualitatively different data bases such as Dialog or BRS. By definition, then, for a data collection to be part of the System VAPORTRAILS must know how to *reach* it via some kind of interactive network protocol.

2.2. Creating a Path

The first thing I use VAPORTRAILS to do is to aid in the construction of a *path* to the information concerning the Entity. If I have not previously used VAPORTRAILS to locate information about the Entity, VAPORTRAILS presents a top level taxonomy of subjects and I must identify the set of classifications that will lead to the Entity. If I have already located information about the Entity in a previous session I will use what VAPORTRAILS already "knows" about what I have done in the manner described in section 6 below. VAPORTRAILS learns from experience, in the sense that it uses the results of previous searches to present me with suggested list of synonyms and related terms in specific data bases.

From the selected classifications, VAPORTRAILS presents a *directory* of information about which data collections on the System might contain something related to the Entity. VAPORTRAILS does this by offering me tools to browse through information about each data collection. Generalized groupings are presented, i.e. these data collections relate to Medicine, these to U.S. Demographics, these to Russian History and Culture, these relate to Sociology, etc. I am also presented with information about the source of the information in each data collection, its time depth, and the way in which it is structured. From this I will select a list of data collections to search and identify a path which branches first to several data collections and then to one or more subjects in each collection.

2.3. Query Formulation

To formulate my query VAPORTRAILS needs to know two classes of information about each data collection. VAPORTRAILS must know *how to ask a question*, i.e. what the query syntax is for that data base. VAPORTRAILS does not trade functionality for simplicity. The full power of the various syntaxes used by the search software on each data base collection is generalized in the VAPORTRAILS search interface. VAPORTRAILS must also know *what it can ask for*, i.e. what terms are contained in the data collection and how they are classified. VAPORTRAILS directs my search to the extent that it does not allow me to ask for terms which are not available.

Given a list of data collections VAPORTRAILS now presents information about the classification structure used by each branch in the path. I am given tools for browsing lists of terms which are indexed, with visual cues to distinguish generally indexed terms (i.e. this term may appear in a title or abstract) from subject classifications. I am given a graphic representation of the level of complexity each data collection uses for classifying their subject information. This may be either a flat list of subject terms or a graphic tree structure. With these tools I must identify the most likely synonyms in each classification structure which

describe the Entity. When the classification system is a tree structure, I need to determine how high up in the tree I want to begin my search.

VAPORTRAILS understands two general kinds of qualitative differences among data bases: those that are functionally equivalent but differ in implementation and design, and those that are functionally different.

There are data bases which are functionally equivalent but differ in the way citations are formed and the subject context in which terms are understood. The System may contain 100 data bases, all of which are bibliographic in nature, ranging over several subject areas of medicine, psychology, sociology, anthropology, biology, religion, etc. The System may also contain 50 different full text data bases of newspaper, newsletters and magazine covering the subjects of electronics, medical research, current events, movie gossip, and the construction industry. The way the material is stored and classified will varies in each of these cases, but VAPORTRAILS will present and retrieve it in a uniform manner.

VAPORTRAILS understands collections of data bases which are functionally different as well. The System may contain a collection of bibliographic data bases on public policy issues, full text collections of government reports, various business and news periodicals, along with numeric information on businesses and regional demographics, along with numerical map data. Having a general model for each of these kinds of data bases, VAPORTRAILS provides user interface tools which are appropriate to each kind of data.

VAPORTRAILS provides a convenient interface for building a complex query. A general subset of searching features are presented when building a query that is *not yet* associated with a specific data collection. When building a query for a specific data collection the interface changes to present the full set of search features specific to that data collection. Having generalized the search syntax for locating information anywhere on the System, VAPORTRAILS knows how to modify a query when searching across different data bases. If the differences are logical exclusions or syntactic translations, VAPORTRAILS will do these without my intervention. If I have used a feature when searching in one data collection which is not available in another, VAPORTRAILS will tell me what I have to change. If the search terms I have used are not available in each data base, VAPORTRAILS will provide information for me to reformulate my query.

I may wish to prepare queries for several data collections at once, or I may wish to start at one branch in the path and refine that based on a set of iterations before proceeding to the next branch. In either case the search of each data collection is done in serial or in parallel when possible. The result of a search in each data collection is treated as a discrete set, to be manipulated and evaluated as described below.

2.4. Relevance Judgment

VAPORTRAILS takes my query and comes back with a set of results from each data collection. These results are presented as a set of graphic icons. The purpose of these icons is to provide me with a simple way to judge the relevance of the results. VAPORTRAILS provides me with tools for adjusting the design and the physical placement of the icons on my display so that I can bring these results into visual focus.

Each icon presents a visual condensation of the information contained in a record (citation, data set, document, image) from that data collection. Selecting the icon presents the information in the record itself. The results of selecting an icon will be appropriate to the data collection, and will vary from a formatted set of fields in a bibliographic citation, to a scrollable view of a text or text and graphics document, to a view of a graphics image.

VAPORTRAILS also provides me with tools for reformulating my query based on my results. VAPORTRAILS collects statistics (bibliometrics) about each set of results, and presents this information both in the design and physical placement of the icons and in the form of lists. Overlaps in subject terms,

unanticipated but interesting subject headings, high frequency of certain authors and sources are presented so that I can easily use this information to refine or alter my original search path.

2.5. Retrieval of Information

The data collections on the System may fall into three access categories: browse-only, browse-and-retrieve, and browse-retrieve-and-alter. Information in a browse-only data base may not be copied to the local system in any manner other than visual presentation on the screen for legal (copyright) or technical (data set is simply to large to transfer) reasons. I can make a local copy of information in the browse-and-retrieve data base, and in the case of the browse-retrieve-and-alter data base I can also choose to modify the original record. In the case of the latter two categories, once I have identified a set of results which relate to the Entity I can choose to retrieve the information in these records, i.e. locally save a copy in VAPORTRAILS.

When retrieving data VAPORTRAILS translates the structure of the original record into its own record structure. VAPORTRAILS will parse the original record, then do its best to atomize all information in the original record into discrete fields. Examples would be breaking up the contents of an author field into separate author names, and breaking up each name into separate parts; breaking up a source field into journal name, volume and issue numbers, and page numbers.

VAPORTRAILS will also assign a *type* to the record for local classification. Examples of types would be bibliographic types such as journal, book, dissertation, conference proceedings; full text of articles or books; tables of data; digitized or vectorized images; directory information.

The name of the original data collection and the subject classification system of the original data base is retained and stored with each record. VAPORTRAILS uses this information for suggesting refinements when I assign indexing terms to the record. This information will be used for suggesting paths to follow in later searches.

VAPORTRAILS provides me with a way of assigning index terms to each record I retrieve. Tools are provided to maintain consistency in other areas which are *not* consistent across qualitatively different data bases such as author names, abbreviation of journal titles, dates, etc. This provides the personal equivalent of library science concept of *authority control*. A *taxonomy processor*, akin to the current outline processor model, is provided to aid in the creation and assignment of controlled vocabulary.

2.6. Refining and Re-using Citations or Documents

Once citations or documents are selected for retrieval they are stored by VAPORTRAILS so that they can be used and re-used in several different ways.

By storing the data base source and the original subject headings with each record VAPORTRAILS is able to build up a body of knowledge about how an Entity has been located. VAPORTRAILS also stores a *search strategy* used to locate the Entity, which may be viewed and re-used. The index key words I have assigned to a record are now synonyms for the original subject heading in the data base source. VAPORTRAILS can apply the search strategy and the index synonyms when I seek a path to this or a related Entity in that data base.

Once a document is retrieved I am able to use VAPORTRAILS to regularize and improve it. Four sub-applications are provided for handling bibliographic data, full-text data, numeric data, and image data.

In the case of bibliographic and full-text data, I can edit it to add to or modify an existing abstract. Since VAPORTRAILS is part of my hypertext system, I can further annotate citations, full text documents, or images by adding links to other citations, documents, or images. I can add local holding information to bibliographic citations, so that VAPORTRAILS can later be used to locate material for reading or copying. I can index and then search full text records to locate and copy material to support my own writing. The

numeric data sub-system can perform statistical analysis and transformations on tables of data, and can present the data to graphics applications for a variety of graphic presentations. I can use VAPORTRAILS to locate citations for proposals, articles, and books. In cooperation with my document preparation system, VAPORTRAILS can produce citations in the form of footnotes and bibliographic lists, formatted to any bibliographic style. VAPORTRAILS can also be used to produce documents which are themselves detailed bibliographies.

The local data base built from retrieved citations, data sets, and documents is itself a major source of information. VAPORTRAILS allows me to browse and search these records with the same interface I use for other data bases. Since I know my own controlled indexing vocabulary, and VAPORTRAILS knows its own structure and syntax, retrieval is simple and precise.

The VAPORTRAILS data base can accept information from retrieved records, from any set of structured records from another data base system, or from keyboard input. This data base is easily extensible, and appropriate for maintaining reading lists for courses, special topic bibliographies, cataloging special collections, and full text indexing.

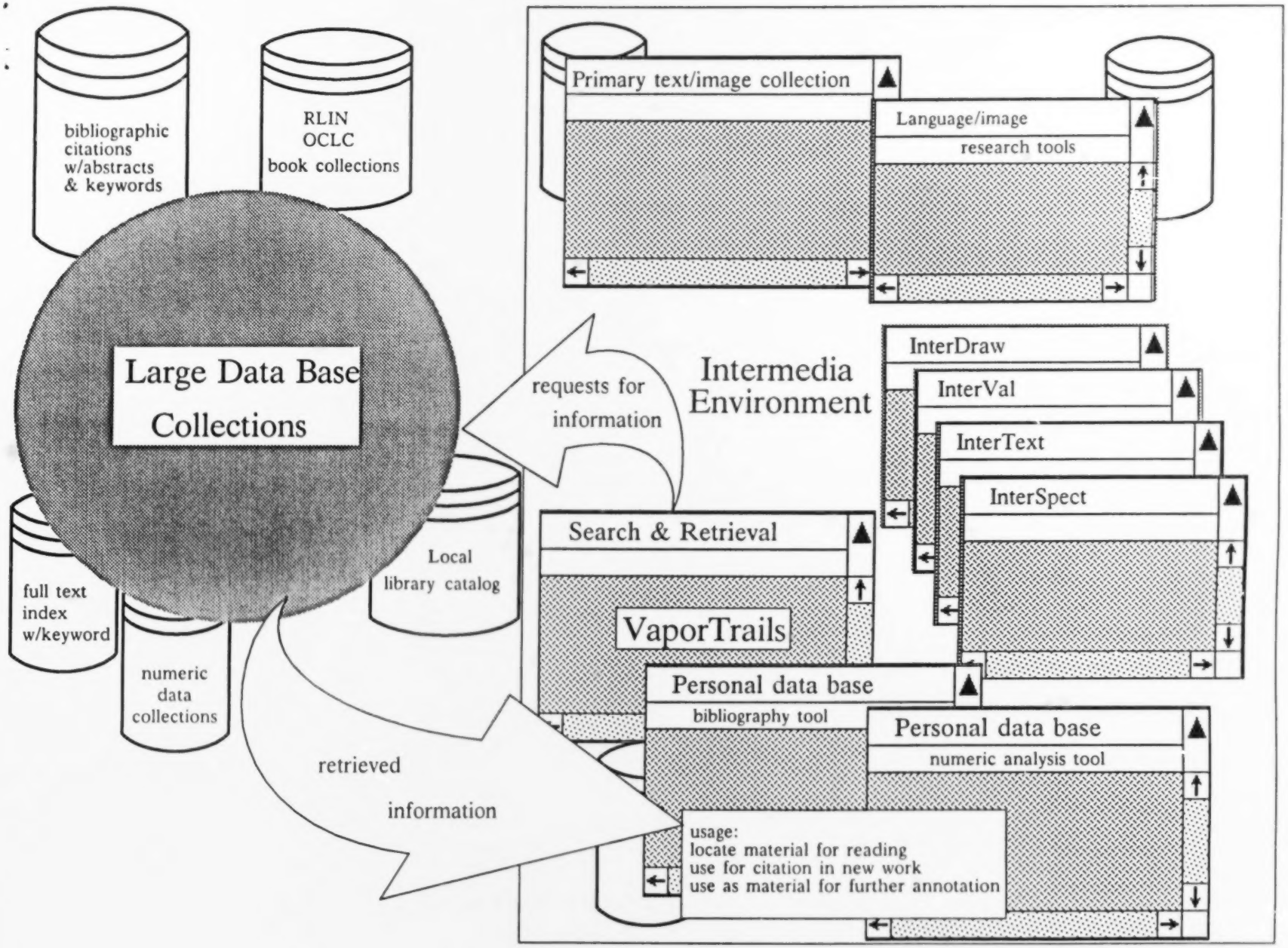
2.7. Customization and Extensibility

As describe above VAPORTRAILS must have specific information about a data collection before it can search in it. It must know how to locate the data collection. It must know how to communicate with the computer on which the data collection resides, and how to interact with the host data base software that accesses information in that data collection. It must have a local copy of the index of the data collection, or a method of accessing that index in a timely and convenient manner.

When a data base is added to the Dialog system, the data base supplier provides Dialog with information about how the data base is structured. Dialog creates and distributes a "blue sheet" to all Dialog subscribers. The subscriber then uses this information when accessing the data collection. This paper analogy is a model for the more detailed *data base spec* that must be supplied to VAPORTRAILS as a structured file. The data base spec, or the information needed to create it, can be created by the data base supplier. Where it is not supplied tt may be gathered by a central administrator for use on many machines. In cases where this information already exists, i.e. where an instance of a data base has been specified for an instance of VAPORTRAILS, it can be copied and loaded from system to another.

VAPORTRAILS contains a set of tools I can use to supply this data base spec myself. Using a series of menus and forms I can create this structured file from written documentation provided by the data base supplier.

These same tools can be used to define and extend my personal data base. As I use it to collect and manage my own information I define the structure I want the data to have. I also use these tools to define bibliographic formats I want to use when extracting information for footnotes and citations and tabular and graphic format I want to use to extract presentations of data sets I have stored. If I choose to share my personal data base with other VAPORTRAILS users I can supply them with my own data base spec so that they may access my data base in the same manner they access other sets of data.


bibliographic citations w/abstracts & keywords
RLIN OCLC book collections
Large Data Base Collections
full text index w/keyword
numeric data collections
Local library catalog
requests for information
retrieved information
Intermedia Environment
Primary text/image collection
Language/image research tools
InterDraw
InterVal
InterText
InterSpect
Search & Retrieval
VaporTrails
Personal data base
bibliography tool
Personal data base
numeric analysis tool
usage:
locate material for reading
use for citation in new work
use as material for further annotation

3. Large Data Bases and Intermedia Extensions

We propose that the scholar needs access to four classes of data resources:

1. *Primary text/data/image collection(s) in the field*: collections of material which are themselves the subject of study.

This is original source material which is the object of scholarly examination. Examples include the Thesaurus Linguae Graecae data bank, all of Shakespeare, all of Dante, all of Jung and Freud, a collection of DNA sequences, LandSat data, econometric data sets. The taxonomy of access to such material must be tailored to the material itself. In the case of text this should be linguistic and structural.

2. *Bibliographic citation and reference material*: collections of information describing original source materials and secondary material.

This kind of information is used for location, identification, or citation purposes. These descriptions currently make up most of the large on-line data bases collections on Dialog and BRS, consisting of bibliographic references indexed by field. Important examples include the MLA Index, Religion Index, RILA, Chemical Abstracts, OCLC book catalog, Sociological Abstracts, ISI Science Citation Index. This bibliographic material is used to lead the scholar to primary and secondary work about a particular subject.

3. *Language/image reference material*: collections of information that are used to analyze or better understand primary material.

These information collections are applied directly to primary material, to assist in the scholar's investigation of that material. Commercial examples include spelling and grammar checking systems accessing dictionary and thesaurus (MacLightning, Microsoft Bookshelf, various versions of the American Heritage Dictionary), subject-specific reference works such as encyclopedias (Grolier Academic American Encyclopedia), glossaries, indexes, map data.

4. *Personal information resources*: collections of annotated bibliographic information and other data selected for personal use, as well as other forms of cross referencing and annotation materials.

These are collections synthesized and *personalized* from other sources by the scholar, specific to the scholar's particular research interests and objectives. Current examples of applications software for manipulating such personal collections of sources include microcomputer-based bibliographic data base management systems such as Professional Bibliographic System, Pro-Cite, Sci-Mate Manager/Editor.

The preceding illustration shows an applications window for each of these classes of data collections contained within the frame of the Intermedia Environment.

The capabilities of Intermedia (as of Spring 1987) address the first and fourth categories described above. Given a body of information stored as text, vector graphics, and raster graphic files, Intermedia allows the user to create, view, and edit webs of association among the material contained in those files by placing links between items. A Link between two Blocks can be named and these names can in turn be used to direct the user's choice of which path to follow when viewing the data.

3.1. Access to Qualitatively Different Bibliographic Citation and Reference Material

In the current Intermedia system all data must be contained in files that are created and/or read by one of the four Intermedia applications. We want to add VAPORTRAILS, a general data base search tool for accessing information located in large central data base collections external to Intermedia from within the Intermedia environment. This application assumes that these collections of data are qualitatively different from each other and from the user's own personal data base, but it maintains a generalized model which is used to describe all the data bases in a uniform manner. We want to present the user with a single

interface paradigm through which these qualitatively different sets of data can be accessed and manipulated.

One paradigm for the user interface model is the Graphic Icon-driven Search Tool (GIST) prototype developed by Page Elmore. In GIST the user is presented with a search form, and a query is developed by selecting terms from the presentation of a controlled vocabulary appropriate to a particular data base. It is assumed that the results will be browsed and used to reformulate the query. The results are presented as graphical icons to aid in the selection and reformulation process.

The application will make use of the record/field structure and the controlled vocabulary for each field in the data base (variously known as the thesaurus, authority records, data dictionary, etc.). One research question is how to best make this kind of information available to the user through the application. Should this information be accessed directly by the user simply as a separate file, following the example of a printed thesaurus or Dialog on-line thesaurus, or should it be a set of data which is *probed* by the application in response to the user's query? The current GIST prototype uses the technique of presenting the user with a controlled vocabulary list from which to build queries. Where this vocabulary comes from is something of a chicken-and-egg problem. Certain academic disciplines have a recognized or de facto standard of thesaurus terms for classification, which may represent a simple or complex tree structure. This may be the best base on which to build the user's own *local* classification system of keywords or index terms. With this model the user would begin with a recognized thesaurus of terms for the field in which he is searching. Using some combination of terms selected from this base plus author/title/text information he retrieves a set of records from a data base. When he identifies a set of retrieved records as acceptable they are processed for storage in the local bibliographic data base. These local copies contain the controlled vocabulary keywords assigned to them from the source data base. The user annotates these records by adding a *local* classification of keywords. The GIST application can then make use of the correlation between these two sets of keywords when making a future query against this same source data base.

For example:

The user wants to look for bibliographic citations about Buddhist influences in British and American literature. There are several data bases which can be searched: MLA Index, OCLC EASI Reference, Religion Index, RLIN, Arts & Humanities Citation Index, Book Review Index. Each has its own unique classification system. VAPORTRAILS first leads the user through the classification systems for each data base, helps the user to build the best query, then executes and retrieves the information. The useful records are stored and indexed by a uniform set of keywords selected by the user. The user builds up a set of citations by and about a group of authors in this manner. When he then goes to look for more information about these authors, not necessarily having to do with the original questions, VAPORTRAILS "knows" how to ask about each author in each data base by referencing the original keyword field in the locally stored records.

The user is interested in finding bibliographic references about social theories of work and how they are applied to work in university environments. He is further interested in how the impact of computers on the workplace has been studied and what if any correlations there are between these studies and the more general social theories. Each of these topics can be searched in ERIC, LISA, Sociological Abstracts, Dissertation Abstracts, etc. VAPORTRAILS leads the user through the classification system for each data base and builds up a knowledge base of how to correlate the user's terminology for each concept with the index terms in each data collection. Sets of synonyms are collected and stored for future use. As each query is performed the user selects records which look promising, and expands the search based on related index terms in each useful reference. VAPORTRAILS calculates bibliometrics on each of these selections, showing the user which authors and journals are the most frequent. Such a profile is applied by the user to continue refining and expanding the search.

3.2. Action Links

In the current model if the user wishes to associate a Block containing of a name in one document with a Block containing a biography in another document he creates a Link. Once this Link is created the selection of the name in the first document allows the user to follow to a biography. We want to add access to Language/image research tool data bases through applications which perform actions based on the process of selecting an object in an Intermedia document. If a geographic/biographic dictionary or glossary in that field is integrated into Intermedia as a research tool, the user would be able to select *any* proper name and select a *follow dictionary* function which would create an *action link* between the string in the original document and an entry in the dictionary data base. The *action* in this case will be a call to a lookup procedure.

Here are several examples:

The Primary data base is the prose text used for teaching English to foreign language students, along with other prose texts which provide practical information about student life at Brown. An American English dictionary, an geographic/biographic lexicon, an English language thesaurus are available as electronic resources. The instructor can link sentences in the prose text to simplified sentences and grammatical examples for the students to follow. Related passages about student life can also be linked. The student can choose to follow the link to a simplified sentence when it is helpful. He can select any word in the prose text and select *follow dictionary* to locate the dictionary article on that word. Texts which explain practical aspects of student life at Brown, such as where the health services are, and how one goes about registering for classes, are built by linking explanatory texts with maps of campus and example registration forms.

The Primary data base is a collection of digitized drawings and photographs, along with supporting text documents related to these works of art. The RILA bibliographic data base of works about the artists and another data base which catalogs the location of works by each artists are available as electronic resources. While viewing a drawing the user can select the name of the artist and select *locate works* to see a list of where the works of this artist can be found. The result can be filtered so the user might only see those works located at one institution, in one country, or world-wide (see filtering below). The user can also select *locate citations* and bring up a data base search application that can be used to locate citations about this artist in RILA. Useful citations from RILA can be copied to a local data base and linked to the web of supporting text documents about the author.

The Primary data base is the collection of Classical Greek texts in the TLG data base and a similar collection of Classical Latin texts. The user wishes to identify which writers make reference to Alexandria, Egypt after the first century A.D. Using a Greek and Latin lexicon application the user identifies the possible names for Alexandria. Using the TLG *Canon* the user creates a list of authors who wrote after the specified date. The full-text index of the text of each of these authors is searched for these terms, and the resulting set of authors who mention Alexandria is presented to the user. The user creates a *document* which consists of the bibliographic references to the work of each author, and a block of text from each work. Browsing this document the user finds a description which is of interest in the work of an author who wrote during the third century A.D. The user selects a map data base which contains information on the Mediterranean basin between eighth century B.C. and seventh century A.D. Using a map browser he selects the Alexandria area at a time depth of third century A.D. The resulting two-dimensional image contains a number of building and site markers for which additional information is available. Selection of a site marker brings up two- and three-dimensional building and site reconstructions based on archaeological evidence. The user may choose to create an Intermedia link between this site plan and the passage in the primary author, which he believes refers to a building on this site at this time depth.

The Primary data base is the collected writings of the twentieth century American poet Ezra Pound. This consists of his collected short poems *Persona* and the long poem *The Cantos*, his prose essays, his translations of poetry from ten different languages, a bibliography of published works by the poet, a bibliography of works about the poet, and selected support materials. The user can go to a particular work by specifying a title or looking for a word or group of words. Any word in any work can be selected and a dictionary application for for that language can be selected, so that *follow dictionary* will locate an English language definition of the word. Any text which is a translation is linked on a document level to that text in the original language, so that the user can select *show original* to display the original text. Selected works are also linked to scanned images of original manuscript versions of the text, so that the user can select *show manuscript* to display and compare a manuscript version and the finished work. The bibliography of secondary works is indexed by work, poem, and concept. The user can be reading a particular poem, select *show bibliography* and see a list of those books and articles which make reference to that particular poem. From the same point the user can annotate a poem or block of text in a poem with an associated keyword, which can later be used to create a *link document* consisting of those blocks of text associated with that keyword.

3.3. Searching in Intermedia

We want to generalize the concept of searching for an Entity whether that Entity is to be found outside or inside the Intermedia framework. We have already suggested that VAPORTRAILS would have tools for creating a personal data base. This would consist of records the user has located outside and then copied into the Intermedia framework. These records could then be manipulated in a number of ways: edited, annotated, etc. It could also consist of records entered directly in VAPORTRAILS. This data base could also be seen as containing information about all the documents within the Intermedia framework itself.

In the case of documents within Intermedia, the user should be able to search for any Block which meets some criteria, in the same way they would search for a citation. Where relevant the type of Link which leads to the Block can be a criteria in the search. A Block, i.e. a defined location in a document, would have associated subject headings (keywords). These keywords would be related to information about the document in which the Block resides, and information about the creation of the Block: the author of the Block, information about any Link made to or from the Block, etc. The user could browse a list of these subject headings, then build a query which would result in the location of those Blocks meeting a specified criteria. The result of the search would be information about the Blocks, analogous to a bibliographic citation. If the user chose to follow this information, Intermedia would respond by opening the document and locating the contents of the Block itself.

The scope of a query can be limited in several ways. The set of documents in which to search for Blocks can be limited by a list of webs, folders, or documents. A search can also be reformulated and limited only the results of a previous search.

Citations located outside the Intermedia framework, citations which refer to Blocks, and Blocks in a document can be linked together in several ways. For example:

A group of students are interested in writing a paper about *The Rape of the Locke* by Alexander Pope. A copy of the poem is created in InterText. Using VAPORTRAILS they search for articles about this poem in the MLA Index, then follow these references to located other works in Arts and Humanities Citation Index. The resulting set of secondary sources is selected and used to create a data base. Local holding information on the cited journals is retrieved from JOSIAH. Once the students have read the material they annotate the references by adding their own abstracts and creating links from the citations to Blocks in the document containing the poem. The English 32 web is also searched for material related to the poem, poet, time period, and genre. The resulting Blocks are browsed. These contain references to other secondary sources and issues which lead the students to further searching on external data bases. Once

the research phase is well underway, the students begins to compose their papers. Where they make reference to secondary sources they create a link to the citation in the data base. Where they quote from the primary text itself they also make a link. Students reading the primary text or browsing the citations can follow links to relevant passages in other student's papers.